GWAS-QTL Colocalization

Annie Novak, Henry Wittich, and Maleeha Fatima

Abstract

In order to understand the underlying molecular mechanisms involved in polygenic diseases, scientists must uncover the genetic variants involved and how they affect phenotype. A method now widely used for this process is colocalizing significant genome-wide association studies (GWAS) and expression quantitative trait loci (eQTL) signals. Tools such as COLOC use statistical tests to determine colocalization and provide potential causal links between single nucleotide polymorphisms (SNPs) that alter gene regulation ultimately contributing to disease traits. One of the necessary inputs to COLOC are linkage disequilibrium (LD) matrices and there are many tools available that provide reference LD matrices; however, these tools mostly use European populations to build them. To facilitate a smoother process for those performing COLOC analysis in admixed populations, this pipeline builds LD matrices specific to the input population, formats GWAS summary statistics and any form of QTL data, and runs COLOC for the user.

Introduction

The majority of common diseases are not caused by a mutation in a single gene but rather by several genetic variants or SNPs, and each one can have a varying effect on phenotype. GWAS are used to identify which SNPs are highly associated with a complex disease trait. Additionally, eQTL mapping finds associations between SNPs and gene expression levels. Since over 90% of significant GWAS variants occur in non-coding regions, SNPs are likely influencing disease traits indirectly through gene regulatory changes[1]. By finding colocalized eQTL and GWAS signals at the same locus, we can establish causality for which SNPs affect gene expression changes that are in turn contributing to a disease trait. However, only visually identifying colocalized signals can lead to false positives. First, an abundance of eQTLs in the human genome increases the probability of an overlap occurring simply by chance[1]. Second, some SNPs may be in LD with the true causal SNP, making them also appear significant. Therefore, tools have been developed to apply statistical tests to determine the probability of colocalization. Tools and methods that have been used for colocalization analysis include the proportionality test[2], COLOC[3], and eCAVIAR[4].

This GWAS-QTL colocalization pipeline provides a stream-lined way to prepare GWAS summary statistics, any type of QTL data (eQTL, pQTL, mQTL), and LD matrices for input and then perform COLOC analysis (implemented online at https://github.com/annie-novak9/Coloc). Most importantly, this tool will build LD matrices specific to any admixed population input by the user. Using the colocalization results output by the pipeline, a user can begin piecing together the molecular mechanisms contributing to a complex trait and identify candidate disease genes for future functional studies.

Implementation

Our pipeline is written in R, bash, and Python. The main packages used are COLOC and PLINK. The pipeline consists of 5 scripts which together create LD matrices as well as format QTL data and GWAS summary statistics to be input into COLOC for analysis. The pipeline requires the following inputs: SNP annotation files, gene annotation file, genotype files, population and chromosome dosage files, GWAS summary statistics, frq file, QTL file, populations, population size, chromosomes, and a gene IDs list (optional). Script one takes the SNP annotation, gene annotation, and genotype files and outputs 1 MB of each gene's transcriptome. Then, script two formats the chromosome dosage files to a binary format using PLINK. The outputs of scripts one and two are then used in script three to create LD matrices using PLINK. Script four ensures that the QTL and GWAS summary statistic files are properly formatted. The output from scripts three and four are then input into script five, which runs COLOC. The pipeline then outputs a COLOC results summary table of each gene, containing p-values for five hypotheses (Table 1). The hypotheses of most relevance are H_3 and H_4. If the p-value for H_4 is significantly greater than H_3, then a user can conclude that the GWAS and QTL signals are colocalized due to a causal SNP.

The functionality we developed in this pipeline is the main wrapper, which consolidates all five scripts to hide their complexities and allows the user to only input the parameters once. The maximum output memory used while running the pipeline and the amount of time it takes was calculated by using running the pipeline with the *time --verbose* command in the command line.

Results and Discussion

In order to test the pipeline, we had to combine data from two populations, Americans of African ancestry in southwest USA (ASW) and African American (AFA), in order to have sufficient information[5,6,7]. The phenotypes we tested are BMI, C-reactive protein levels, HDL cholesterol, and total cholesterol. Table 1 shows a subset of the COLOC results output from the pipeline using our test data for BMI. The genes highlighted in green are significantly colocalized based on the comparison of their H_3 and H_4 p-values; meanwhile, those highlighted in red are not.

Table 1. Pipeline output of COLOC results summary for ASW/AFA populations for BMI.

hit2	hit1	nsnps	PP.H0.abf	PP.H1.abf	PP.H2.abf	PP.H3.abf	PP.H4.abf	best1	best2	best4	hit1.margz	hit2.margz	gene
chr3:53187662	chr3:52286211	5	0.997701	0.001703	0.000445	7.60E-07	0.000150252	chr3:52286211	chr3:53187662	chr3:52286211	-2.932	-0.85986	SL000554_ENSG00000163932.15
chr1:169102398	chr1:169102398	9	0.998932	0.000193	0.000855	1.66E-07	1.94E-05	chr1:169102398	chr1:169102398	chr1:169102398	-1.43693	1.674421	SL000560_ENSG00000174175.17
chr1:204307264	chr1:203553781	7	0.999032	0.000238	0.000707	1.69E-07	2.22E-05	chr1:203553781	chr1:204307264	chr1:203553781	2.348315	-2.30434	SL000565_ENSG00000143839.15
chr1:159031275	chr1:158875526	8	0.998777	0.000284	0.000911	2.59E-07	2.84E-05	chr1:160112660	chr1:159031275	chr1:160112660	1.545832	-1.88973	SL000573_ENSG00000132703.4
chr2:482688	chr2:621954	1828	4.92E-08	0.794412	8.23E-09	0.133048551	0.072539288	chr2:621954	chr2:490646	chr2:621954	7.095983	2.98983	SL000588_ENSG00000115705.21
chr2:482688	chr2:633980	1828	0	0.799082	0	0.133830727	0.067086944	chr2:633980	chr2:490646	chr2:633980	24725.54	2.98983	SL000588_ENSG00000115705.21

Several tools exist that perform statistical tests for colocalization. Older methods like the proportionality test are limited in that they require individual genotypes as input data and don't consider LD[2]. Newer methods like COLOC and eCAVIAR confront these limitations by incorporating LD matrices into their analysis and allowing GWAS summary statistics data as input[3,4]. Our implementation of COLOC is novel because it builds LD matrices for any input population. This makes our pipeline suitable for diverse and admixed populations that do not have pre-made LD resources available.

The maximum output memory used while running the pipeline for each chromosome we tested is shown in Figure 1. The size did not vary for different numbers of phenotypes. The time it takes to run the pipeline for each chromosome with varying numbers of phenotypes is presented in Figure 2.

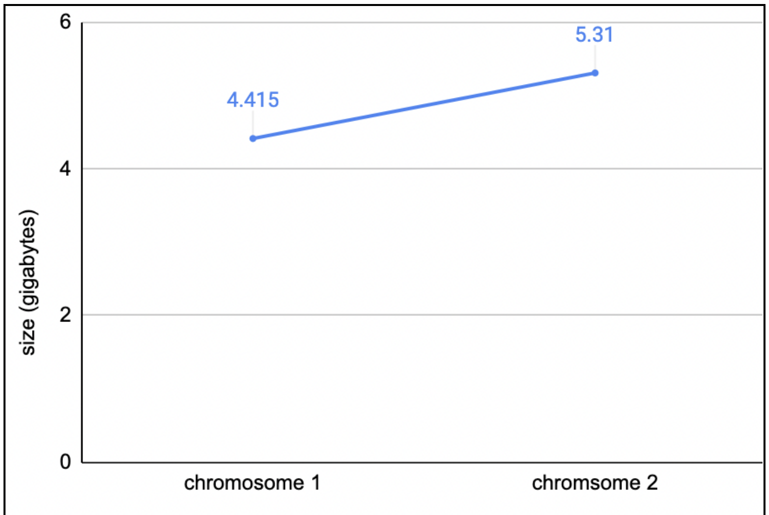

Figure 1: The maximum output memory usage for each chromosome.

One improvement we could implement in the future is to add in a parallelization functionality so that the pipeline can run each chromosome at the same time rather than looping through each one. We could also incorporate the S-PrediXcan script back into the pipeline so that it will only run COLOC on significant genes output from S-PrediXcan. In addition, our pipeline currently outputs a summary of all the COLOC results without any further analysis. This can be optimized by analyzing the COLOC results to only output genes that are significantly colocalized. This could be done by filtering the COLOC results for genes that have a significantly greater H_4 than H_3 p-value.

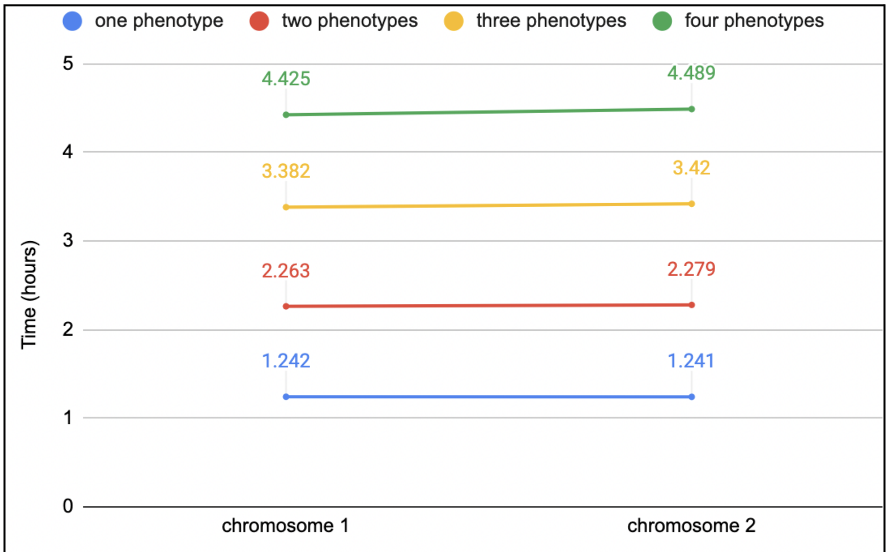

Figure 2: pipeline running time for each chromosome with varying numbers of phenotypes

References

1. Cano-Gamez E, Trynka G. From GWAS to Function: Using Functional Genomics to Identify the Mechanisms Underlying Complex Diseases. Front Genet. 2020 May 13;11:424. doi: 10.3389/fgene.2020.00424.

2. Wallace C. Statistical Testing of Shared Genetic Control for Potentially Related Traits. Genet Epidemiol. 2013 Nov 5;37(8): 802-813. doi: 10.1002/gepi.21765

3. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et. al. Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics. PLoS Genet. 2014 May 14;10(5): e1004383. doi: 10.1371/journal.pgen.1004383.

4. Hormozdiari F, Bunt MVD, Segré AV, Li X, Joo JWJ, Bilow M, et. al. Colocalization of GWAS and eQTL Signals Detects Target Genes. 2016 Dec 1;99(6): 1245-1260. doi: 10.1016/j.ajhg.2016.10.003

5. 1000 Genomes Project Consortium, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68-74. doi: 10.1038/nature15393. PMID: 26432245; PMCID: PMC4750478.

6. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, Greenland P, Jacob DR Jr, Kronmal R, Liu K, Nelson JC, O'Leary D, Saad MF, Shea S, Szklo M, Tracy RP. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002 Nov 1;156(9):871-81. doi: 10.1093/aje/kwf113.

7. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, Howard TD, Hawkins GA, Cui W, Morris J, Smith SG, Barr RG, Kaufman JD, Burke GL, Post W, Shea S, McCall CE, Siscovick D, Jacobs DR Jr, Tracy RP, Herrington DM, Hoeschele I. Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013 Dec 15;22(24):5065-74. doi: 10.1093/hmg/ddt356.